Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 45 (2015) 86 – 94

Procedia Computer Science

International Conference on Advanced Computing Technologies and Applications (ICACTA-2015)

Identifying Concept-Drift in Twitter Streams

Lifna C.S, Dr. Vijayalakshmi M *

Information Technology, VESIT, Mumbai – 400074, India

Abstract

We live in a Big Data society, where the dignity of data is like exchange of currency. What we produce as data affords as access to different application, benefits, services, delivery etc… In today's world communication is mainly through social networking sites like, Twitter, Facebook, and Google+. Huge amount of data that is being generated and shared across these micro-blogging sites, serves as a good source of Big Data Streams for analysis. As the topic of discussion changes drastically, the relevance of data is temporal, which leads to concept-drift. Identification and handling of this concept-drift in such Big Data Streams is present area of interest. The state-of-the-art techniques for identifying trending topics in such data streams mainly concentrates on the frequency of the topic as the key parameter. Concentrating on such a weak indicator, reduces the precision of mining. This study puts forward a novel approach towards identifying concept-drift by initially grouping topics into classes and assigning weight-age for each class, using sliding window processing model upon Twitter streams.

Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).

Keywords: Big Data; Data Stream Mining;Sliding Window; Landmark Window; Twitter; Classification; Topic ranking; Concept-Drift.

1. Introduction

In the digital world, things that are repeated become reputable or trendy. Our different practice of liking, sharing, commenting are judged not on the basis of the content, but on the repetition of the content. This might explain why the more popular a tweet is; the more popular it becomes (increases popularity of the tweet). For instance, the proprietary Topic Ranking Algorithm in Twitter, decides what will show up in the coveted real estate of the Top 10

* Corresponding author. Tel.: +91-9969268243; fax: +91-022-61532532.
*E-mail address:*lifnajos2006@gmail.com; viji.murli@gmail.com

Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).
doi:10.1016/j.procs.2015.03.093

Trending Topics on the index page. This depends entirely on the relevance of the topic at that particular instance. The topic keeps changing its position or stick to its previous position as time advances. The topic sometimes slides up or down the rank as it gains public interest or vice-verse. Twitter was therefore selected as a case of study for Concept-Drift Analysis.

In short, Concept-Drift Analysis is the integrated study of identifying and handling Concept-Drift in this evolving stream of data. Adaptation to concept-drift has many real world application such as in Power Industry for Utility Analytics; Telecommunication Industry for Targeted Marketing etc... Our study is an humble attempt to put forth a better algorithm for mining concept-drift in Twitter Streams. Our initial target was to identify concept-drift in such deluge of tweets flowing from Twitter. The remaining part of the paper is organized as follows. Section 2 briefs the state-of-the-art techniques for Topic Ranking and situate our work in the context; Section 3 presents the architectural details of the proposed Concept-Drift Identification System; Section 4 describes the experimental set-up in detail; Section 5 demonstrates the experimental results, and Section 6 concludes with a discussion upon the results obtained and throws light upon the future work.

2. Related Work

In [1] Wang, has come up with a framework, to study concept-drift in political and biological ontologies based on concept identity and morphing. The meaning of concept in these different application domains changes within the given application context, with respect to time. Crammer [2] describes a new topic ranking algorithm (MMP) for multi-labeled documents suitable for massive datasets. The above mentioned state-of-the-art algorithms were experimented in a static environment.

Wang et.al [3], automates on-line news ranking, based on aging theory and burstiness of terms. The topic detection and tracking algorithm also integrates media focus and user attention parameters while ranking. Aiello et.al [4], proposes a topic detection method based on co-occurrence of n-grams and time-dependent ranking after clustering the tweets using LDA. Beel et.al [5], explains the ranking algorithm used in Google Scholar search engine. The paper proposes nine key factors to be considered by any search engine while ranking research papers. Becker [6] proposes a weighted majority algorithm, to rank electrical feeders based on their susceptibility to failure with real-time time-varying data gathered from electricity distribution system. As per the survey performed by Gama [7], concept-drift adaptation technique adopted in one application domain varies from the one adopted for another. Hence for each application domain we need specialized methods to extract insights from the real-time data streams.

Before diving deep into the current research work, it is imperative to have some knowledge of Twitter. Twitter [8] is undoubtedly one of the popular micro-blogging sites [9,10,11] (approx. 500 million tweets sent per day) where users search socio-temporal information [12] such as breaking news, tweets about celebrities, trending topics etc... It is also used as a medium for real-time information dissemination even during election campaigns. Within the Twitter Developers Community there exists many Real-Time Twitter Analytics & Visualization Tools [13], which crawl through Twitter streams by aggregating, slicing, dicing and ranking the data to deliver some meaningful insights on Twitter activities and trends. But all of these consider parameters such as topic frequency or updated frequency as the basic parameter for ranking topics and further analysis.

Our approach in this current work is to rank trending topic by first classifying them and then assigning weight-age for each class based on twelve parameters derived from the top 10 trending topics extracted from Twitter during 2014 Lok Sabha Elections in India.

3. Proposed Concept-Drift Identification System

The proposed Concept-Drift identification System works in four stages as depicted in Fig.1. Overall process can be summarized as follows: Data Collection is the first step, followed by preprocessing of the gathered topics;

Second step is to label the topics into four main classes; in the third step class weight-age is calculated for these labeled trending topics by identifying the twelve dynamic class parameters arising from sliding window processing model. Finally, a graph is plotted for identifying concept-drift in each class based on their weight-age, for a duration of one month.

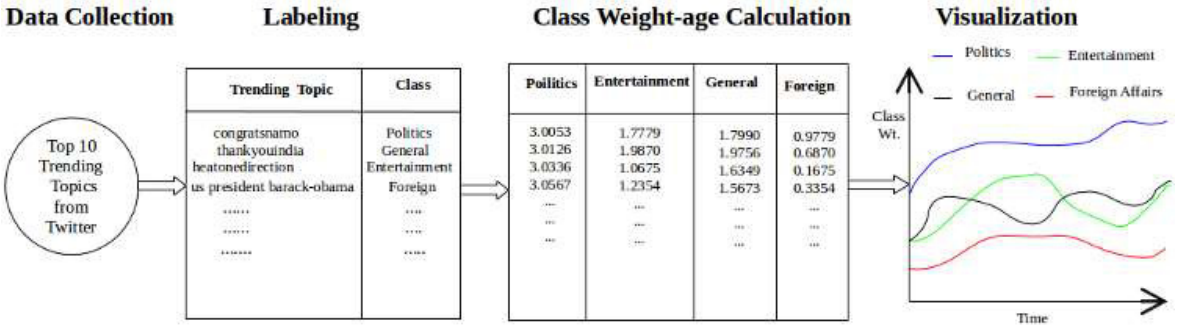

Fig. 1. Proposed Concept-Drift Identification System

3.1. Data Collection

Data Collection was performed over a period of one month which contributed to around 30,000 records as Twitter limits the API requests (15 requests per minute) [14]. Each record contains Top 10 Trending Topics in Twitter which are updated on the website on a real-time basis. Refer Appendix A for raw data downloaded from Twitter. The data is then pre-processed as follows: (1) removing special symbols, (2) inserting hyphens between words, (3) position was appended to each topic, (4) all characters were converted into lower case. And finally a time-stamp generated based on tweet creation time was appended to each record. For the sample data-set refer Appendix B.

3.2. Labeling

Among the trending topics, unique topics were identified and classified into 4 broad categories as Politics, Entertainment, Foreign Affairs, and General. The reason for this classification is to reduce the anomalies which arise due to abrupt changes in the trend. So we confined our analysis upon the classified trending topic data set. Refer Appendix C and Appendix D for category mapping adopted for classification process and sample data-set. As the main thrust of this study was on coining a ranking algorithm, trending topics were classified manually.

3.3. Sliding Window Processing Model for Class Weight-age Calculation

Sliding window processing model was applied upon modified data-set where the window size was set to 10, 20, 30… 100. Weight-age for each class was calculated based on the following twelve Class Parameters identified as shown in Table 1. Five out of these twelve Class Parameters, were identified as Key Parameters for calculating Class weight-age by assigning a weight-age for each parameter determined empirically and normalized as shown in the Table 2. The formula for calculating Class Weight-age is as follows:

$$\textit{Class Weight-age} : \frac{\sum (\textit{parameter\_wt} * \textit{parameter\_val})}{\sum \textit{parameter\_wt}} \quad (1)$$

Table 1. Class Parameters Identified

No	Parameter	Description
1	Class Frequency	Incremented each time the class appears in the window irrespective of its continuity.
2	Class Position Frequency	Incremented with respective to the position of class in the dataset.
3	Previous Class Occurrence	Set to TRUE by default. Set to FALSE, when the class disappears from the dataset in the current record.
4	Chunk Count	Keeps track of the number of chunks in which the class appears.
5	Current Chunk Duration	Keeps track of the duration of the class in each chunk within the window.
6	Updated Class Frequency	Incremented by 1, when class appears in the current record. Decremented by 1, when class disappears till its value is greater than zero and retained as it is when the value becomes zero.
7	Updated Class Relevance	$= \frac{\text{updated Class Frequency}}{\text{window Size (=10)}}$
8	Class Position Weight-age	$= \frac{\sum (\text{pos\_wt} * \text{pos\_Frequency})}{\sum \text{pos\_wt}}$
9	Chunk Start Time Stamp	Records the timestamp of the starting record when the chunk starts or restarts i.e when the class appears or reappears.
10	Previous Chunk Duration	When the class disappear the previous recorded chunk Duration is added. i.e $\sum$ chunk Duration.
11	Class Duration Weight-age	$= \frac{\text{Current Chunk Duration + Previous Chunk Duration}}{\text{window Duration}}$ where, window Duration is calculated as the average window Duration for the dataset considered.
12	Class Relevance	$= \frac{\text{Class Frequency}}{\text{window Size (=10)}}$

Table 2. Five Key Class Parameters for calculating Class Weight-age

No	Key Parameter	Parameter Weight	Maximum Value
1	Previous Class Occurrence	50	TRUE (=1)
2	Class Duration Weight-age	40	1
3	Class Position Weight-age	3	10
4	Updated Class Relevance	2	10
5	Class Relevance	1	10

3.4. Visualization

For quicker analysis of concept drift, graph was plotted using JfreeChart [15] over the duration of the study. Since the study was performed during the 2014 Lok Sabha Elections in India, the Politics-class showed more weight-age throughout the period, followed by General and Entertainment Class. The least weight-age was assigned to Foreign Affairs during the Election Period.

4. Experimental Setup

The experiment was performed on a Intel® Core™ i3-2350M CPU @ 2.30GHz × 4 with 4GB main memory running on Ubuntu 12.04 LTS and all programs were coded in Java using Netbeans IDE 8.0. An application named "IndiaTrend" was created in Twitter for generating API keys and Access Tokens which were used to extract real-time Indian trends from Twitter by setting WOEID = 23424848 via Twitter4J API [16]. The data was collected during 2014 Indian Election process and Java program that was executed on a real time basis program to extract the changing behavior of concepts using sliding window as the processing model.

5. Result Analysis

The experiment reveals that, the identification of concept-drift on trending topics results in ambiguous graphs as shown in Fig. 2. The reason for this behavior is that, trending topics fade and become irrelevant after a period of time.

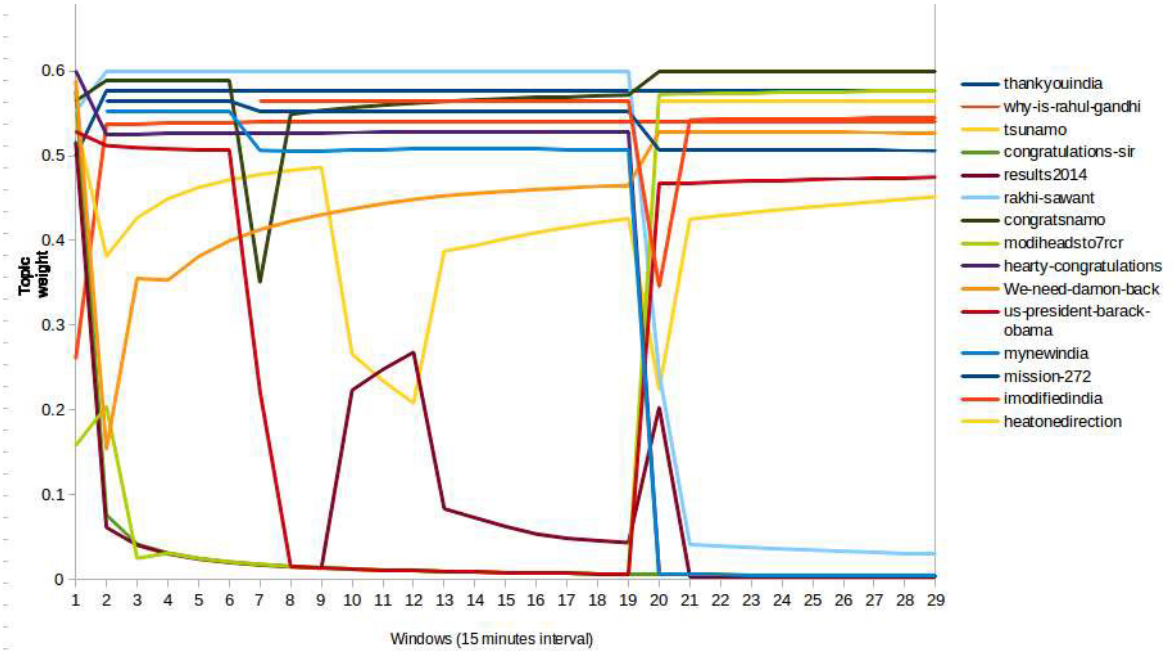

Fig. 2. Concept-Drift in Twitter Trending Topics (window-size = 15)

Hence, the topics needs to be first classified to depict the concept-drift. Next step was to finalize the processing model. Both sliding window and landmark window was applied on the processed Twitter dataset. During the study, it was observed that, the accuracy of class ranking was decreasing in landmark window due to the accumulated class weight-age. So we selected sliding window over landmark processing model.

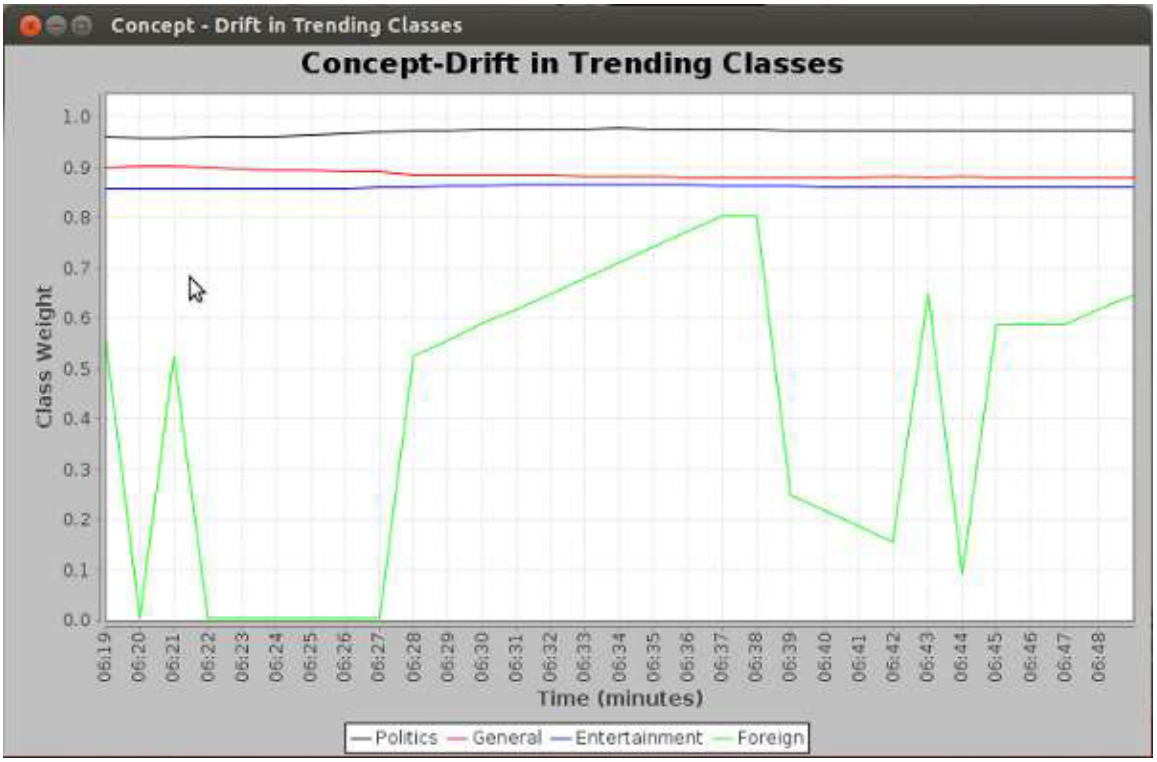

Fig. 3. Concept-Drift in a Classified Trending Topics (window-size = 10)

As a part of result analysis, graphs were plotted by varying window sizes (10, 20, 30… 100). Appendix E depicts the relation between various window sizes and the maximum class weight-age taken by each. Among the various window sizes, window = 10 showed better result as shown in Fig .3. The results showed that, Politics-class gained top ranking in the graph running far above the General, Entertainment and Foreign Affairs category. Of which the top ranking were assigned for the following topics under Politics-class: congratsnamo, modiheadsto7rcr, tsunamo, mission 272. At the declaration of results, trending topics related to Politics-class were at its peak. This clearly reveals the prominence of Politics-class during the period of study.

6. Conclusion

To start with, the dataset was thoroughly studied keeping landmark window as the processing model. After rigorous tests, it was revealed that the model was not ideal for the studying twitter dataset. As in landmark window, temporal span covers all data elements between the starting timestamp to the current timestamp which reduced the relevance of a trending topic class weight-age. Sliding window was selected as an alternative for processing the dataset. Unlike other news datasets, twitter data was undergoing drastic changes and almost no topic was significant over a period of time. Therefore classification needs to be introduced for further analysis. Since the topics were downloaded during 2014 Lok Sabha Elections in India, Politics-class was overruling all the remaining classes. Thus the identification of concept drift from such a turbulent data stream rendered better results.

The study can be extended as an application for real-time analysis of concept-drift over tweets, which can further be mined to extract interesting facts. The classification in this paper was done manually which can be further automated as a future work. Over and above we can improve, the performance of the identification system by incorporating Big Data Tools.

Appendix A. Raw Trending Topics downloaded from Twitter

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
congratsnamo	modiheadsto7rcr	thankyouindia	heatonedirection	imodifiedindia	why is rahul gandhi	we need damon back	mission 272	tsunamo	us president barack obama
congratsnamo	modiheadsto7rcr	thankyouindia	heatonedirection	imodifiedindia	why is rahul gandhi	we need damon back	mission 272	tsunamo	us president barack obama
congratsnamo	modiheadsto7rcr	thankyouindia	heatonedirection	imodifiedindia	congratulations sir	why is rahul gandhi	we need damon back	mission 272	tsunamo
congratsnamo	modiheadsto7rcr	thankyouindia	heatonedirection	imodifiedindia	why is rahul gandhi	we need damon back	mission 272	tsunamo	us president barack obama
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

Appendix B. Trending Topics with position and time-stamp inserted

Time stamp	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1400267418000	congratsnamo_1	modiheadsto7rcr_2	thankyouindia_3	heatonedirection_4	imodifiedindia_5	why-is-rahul-gandhi_6	we-need-damon-back_7	mission-272_8	tsunamo_9	us-president-barack-obama_10
1400267482000	congratsnamo_1	modiheadsto7rcr_2	thankyouindia_3	heatonedirection_4	imodifiedindia_5	why-is-rahul-gandhi_6	we-need-damon-back_7	mission-272_8	tsunamo_9	us-president-barack-obama_10
1400267543000	congratsnamo_1	modiheadsto7rcr_2	thankyouindia_3	heatonedirection_4	imodifiedindia_5	congratulations-sir_6	why-is-rahul-gandhi_7	we-need-damon-back_8	mission-272_9	tsunamo_10
1400267605000	congratsnamo_1	modiheadsto7rcr_2	thankyouindia_3	heatonedirection_4	imodifiedindia_5	why-is-rahul-gandhi_6	we-need-damon-back_7	mission-272_8	tsunamo_9	us-president-barack-obama_10
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

Appendix C. Mapping of Trending Topics into identified classes

Politics	General	Entertainment	Foreign
congratsnamo	thankyouindia	heatonedirection	us president barack obama
modiheadsto7rcr	imodifiedindia	we need damon back	
why is rahul gandhi	congratulations sir	rakhi sawant	
mission 272	hearty congratulations		
tsunamo	mynewindia		
. . .	. . .		

Appendix D. Dataset after Classification

Time stamp	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1400267418000	politics_1	politics_2	general _3	entertain ment_4	general _5	politics _6	entertain ment_7	politics_ 8	politics _9	foreign_ 10
1400267482000	politics_1	politics_2	general _3	entertain ment_4	general _5	politics _6	entertain ment_7	politics_ 8	politics _9	foreign_ 10
1400267543000	politics_1	politics_2	general _3	entertain ment_4	general _5	general _6	politics_ 7	entertain ment_8	politics _9	politics_ 10
1400267605000	politics_1	politics_2	general _3	entertain ment_4	general _5	politics _6	entertain ment_7	politics_ 8	politics _9	foreign_ 10
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

Appendix E. Window Size versus Maximum Class Weight-age

No	Window Size	Maximum Value
1	10	0.9979
2	20	1.4563
3	30	1.9226
4	40	2.3901
5	50	2.854
6	60	3.3197
7	70	3.7863
8	80	4.2532
9	90	4.7229
10	100	5.1882

References

1. Van der Wang, Shenghui, Stefan Schlobach, and Michel Klein. "Concept drift and how to identify it." Web Semantics: Science, Services andAgents on the World Wide Web Vol. 9 Issue..3 (September 2011) Pg. 247-265.
2. Crammer, Koby, and Yoram Singer. "A family of additive online algorithms for category ranking." *The Journal of Machine Learning Research* Vol. 3, Pages. 1025-1058 (2003).
3. Wang, Canhui, et al. "Automatic online news topic ranking using media focus and user attention based on aging theory." P*roceedings of the 17th ACM conference on Information and knowledge management*. ACM, 2008.
4. Aiello, Luca Maria, et al. "Sensing trending topics in Twitter." IEEE Trans. Multimedia Vol. 15, Issue. 6, Pages. 1268-1282 (2013). DOI : 10.1109/TMM.2013.2265080
5. Jöran Beel and Bela Gipp. Google Scholar‘s Ranking Algorithm: An Introductory Overview. Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI’09), volume 1, pages 230 – 241, Rio de Janeiro (Brazil), July 2009.
6. Becker, Hila, and Marta Arias. "Real-time ranking with concept drift using expert advice." *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2007.
7. Gama, João, et al. "A survey on concept drift adaptation." *ACM Computing Surveys (CSUR)* Volume. 46, Issue. 4 (April 2014) Article. 44.
8. Twitter. [Online]. Available : https://twitter.com
9. Alexa Internet, Inc. (2014 December 09). The top 500 sites on the Web. [Online]. Available : http://www.alexa.com/topsites
10. Craig Smith. (2014, October 29). By The Numbers : 250 Amazing Twitter Statistics. [Online]. Available : http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/

11. Thomas Timely. (2014, November 26). Top 10 Micro-blogging Sites. [Online]. Available : http://www.gurugrounds.com/uncategorized/top-10-microblogging-sites/
12. Twitter Milestones. (2014 December 09). [Online]. Available : https://about.twitter.com/milestones
13. Garin Kilpatrick. (2014 November 27). 10 Awesome Twitter Analytics and Visualization Tools. [Online]. Available : http://twittertoolsbook.com/10-awesome-twitter-analytics-visualization-tools/
14. Twitter Streaming API. (2014 December 09). [Online]. Available : https://dev.twitter.com/streaming/overview
15. JFreeChart. (2014 December 09). [Online]. Available : http://www.jfree.org/jfreechart/
16. Twitter4J. (2014 December 09). [Online]. Available : http://twitter4j.org/en/index.html